A Note on: 'Algorithms for Connected Set Cover Problem and Fault-Tolerant Connected Set Cover Problem'

Wei Ren[a], Qing Zhao[a,*]

[a]*Dept. of Electrical and Computer Engineering, University of California, Davis, CA 95616*

Abstract

A flaw in the greedy approximation algorithm proposed by Zhang *et al.* for minimum connected set cover problem is corrected, and a stronger result on the approximation ratio of the modified greedy algorithm is established. The results are now consistent with the existing results on connected dominating set problem which is a special case of the minimum connected set cover problem.

Keywords: connected set cover, greedy algorithm, approximation ratio.

1. Introduction

Let V be a set with a finite number of elements, and $\mathcal{S} = \{S_i \subseteq V : i = 1, ..., n\}$ a collection of subsets of V. Let G be a connected graph with the vertex set $\mathcal{S}$. A *connected set cover* (CSC) $\mathcal{R}$ with respect to $(V, \mathcal{S}, G)$ is a set cover of V such that $\mathcal{R}$ induces a connected subgraph of G. The minimum connected set cover (MCSC) problem is to find a CSC with the minimum number of subsets in $\mathcal{S}$. In [1], Zhang *et al.* proposed a greedy approximation algorithm (Algorithm 2 in [1]) for minimum connected set cover (MCSC) problem, and obtained the approximation ratio of this algorithm. This algorithm has a flaw, and the approximation ratio is incorrect. In this note, we modify the greedy algorithm to fix the flaw and establish the approximation ratio of the modified algorithm. The approximation ratio is with respect to the optimal solution to the set cover problem $(V, \mathcal{S})$, instead of the optimal solution to the MCSC problem $(V, \mathcal{S}, G)$, and thus it is stronger than the one obtained in [1].

2. Greedy Algorithm

Before stating the algorithm, we introduce the following notations and definitions. Most of them have also been used in [1]. For two sets $S_1, S_2 \in \mathcal{S}$, let $\text{dist}_G(S_1, S_2)$ be the length of the shortest path between S_1 and S_2 in the auxiliary graph G, where the length of a path is given by the number of edges; S_1 and S_2 are said to be *graph-adjacent* if they are connected via an edge in G ($i.e.$, $\text{dist}_G(S_1, S_2) = 1$), and they are

*Corresponding author. Phone: 1-530-752-7390. Fax: 1-530-752-8428. Email: qzhao@ucdavis.edu

[0]This work was supported by the Army Research Laboratory NS-CTA under Grant W911NF-09-2-0053.

said to be *cover-adjacent* if $S_1 \cap S_2 \neq \emptyset$. Notice that in general, there is no connection between these two types of adjacency. The *cover-diameter* $D_c(G)$ is defined as the maximum distance between any two cover-adjacent sets, *i.e.,*

$$D_c(G) = \max\{\text{dist}_G(S_1, S_2) \mid S_1, S_2 \in \mathcal{S} \text{ and } S_1, S_2 \text{ are cover-adjacent}\}.$$

At each step of the algorithm, let $\mathcal{R}$ denote the collection of the subsets that have been selected , and U the set of elements of V that have been covered. Given $\mathcal{R} \neq \emptyset$ and a set $S \in \mathcal{S} \setminus \mathcal{R}$, an $\mathcal{R} \to S$ path is a path $\{S_0, S_1, ..., S_k\}$ in G such that (i) $S_0 \in \mathcal{R}$; (ii) $S_k = S$; (iii) $S_1, ..., S_k \in \mathcal{S} \setminus \mathcal{R}$. Let $|P_S|$ denote the length of an $\mathcal{R} \to S$ path P_S, and it is equal to the number of vertices of P_S that does not belong to $\mathcal{R}$. Then we define the weight ratio $e(P_S)$ of P_S as

$$e(P_S) = \frac{|P_S|}{|C(P_S)|}, \tag{1}$$

where $|C(P_S)|$ is the number of elements that are covered by P_S but not covered by $\mathcal{R}$.

For the greedy algorithm in [1], after the subset with the maximum size is selected at the first step, only the subsets that are not in $\mathcal{R}$ and are cover-adjacent with some subset in $\mathcal{R}$ are considered in the following iterations. At some iteration, there may not exist a subset $S \in \mathcal{S} \setminus \mathcal{R}$ that is cover-adjacent to a subset in $\mathcal{R}$, and if we only consider cover-adjacent subsets, then the algorithm will enter a deadlock. Consider a simple example where $V = \{1, 2, 3, 4\}$, $\mathcal{S} = \{\{1, 2\}, \{1\}, \{2\}, \{2, 3\}, \{4\}\}$, and G is a complete graph. If we apply the greedy algorithm in [1] to this MCSC problem, then after $\{1, 2\}$ and $\{2, 3\}$ are selected, the algorithm enters a deadlock.

To fix this problem, we modify the greedy algorithm to include not only cover-adjacent subsets but also graph-adjacent subsets. The modified greedy algorithm for the MCSC problem is presented below.

Input: $(V, \mathcal{S}, G)$.
Output: A connected set cover $\mathcal{R}$.

1. Choose $S_0 \in \mathcal{S}$ such that $|S_0|$ is the maximum, and let $\mathcal{R} = \{S_0\}$ and $U = S_0$.
2. **While** $V \setminus U \neq \emptyset$ **DO**
   2.1. For each $S \in \mathcal{S} \setminus \mathcal{R}$ which is cover-adjacent *or graph-adjacent* with a set in $\mathcal{R}$, find a shortest $\mathcal{R} \to S$ path P_S.
   2.2. Select P_S with the minimum weight ratio $e(P_S)$ defined in (1), and let $\mathcal{R} = \mathcal{R} \cup P_S$ (add all the subsets of P_S to $\mathcal{R}$) and $U = U \cup C(P_S)$.

   End while
3. **Return** $\mathcal{R}$.

3. Approximation Ratio

In [1], the approximation ratio of the greedy algorithm is shown to be $1 + D_C(G) \cdot H(\gamma - 1)$, where $\gamma = \max\{|S| \mid S \in \mathcal{S}\}$ is the maximum size of all the subsets in

$\mathcal{S}$ and $H(\cdot)$ is the harmonic function. In the proof, the authors assume that for every subset S^* in the optimal solution $\mathcal{R}_C^*$ to the MCSC problem, at least one of its elements is covered by the subset S_0 selected by the greedy algorithm at step 1. In general, some S^* may not share any common elements with S_0. Thus, this assumption is invalid, and the resulting approximation ratio is incorrect. In the following theorem, we establish the approximation ratio of the modified greedy algorithm for the MCSC problem. The proof of this theorem does not require this assumption, and it takes into account the additional search of graph-adjacent subsets in the modified algorithm. Furthermore, a stronger result on the approximation ratio is shown in the proof (see Lemma 1). Specifically, the approximation ratio is between the solution returned by the algorithm and the optimal solution to the set cover problem, and the latter is always not greater than the optimal solution to the MCSC problem.

Theorem 1. *Given an MCSC probelm $(V, \mathcal{S}, G)$, the approximation ratio of the modified greedy algorithm is at most $D_C(G)(1 + H(\gamma - 1))$, where $\gamma = \max\{|S| \mid S \in \mathcal{S}\}$ is the maximum size of the subsets in $\mathcal{S}$ and $H(\cdot)$ is the harmonic function.*

PROOF. We show a lemma stronger than the above theorem.

Lemma 1. *Let $\mathcal{R}^*$ be an optimal solution to the set cover problem $\{V, \mathcal{S}\}$, and $\mathcal{R}$ returned by the modified greedy algorithm for the MCSC problem $(V, \mathcal{S}, G)$. Then we have that*

$$\frac{|\mathcal{R}|}{|\mathcal{R}^*|} \leq D_C(G)(1 + H(\gamma - 1)).$$

Let $\mathcal{R}_C^*$ be an optimal solution to the MCSC problem $(V, \mathcal{R}, G)$. Since $|\mathcal{R}^*| \leq |\mathcal{R}_C^*|$, Theorem 1 follows from Lemma 1.

PROOF OF LEMMA 1. The proof is based on the classic charge argument. Each time a subset S_0 (at step 1) or a shortest $\mathcal{R} \to S$ path P_S^* (at step 2) is selected to be added to $\mathcal{R}$, we charge each of the newly covered elements $\frac{1}{|S_0|}$ (at step 1) or $e(P_S^*)$ defined in (1) (at step 2). During the entire procedure, each element of V is charged exactly once. Assume that step 2 is completed in $K - 1$ iterations. Let P_{Si}^* be the shortest $\mathcal{R} \to S$ path selected by the algorithm at iteration i. Let $w(a)$ denote the charge of an element v in V. Then we have

$$\sum_{v \in V} w(v) = \sum_{i=0}^{K-1} \sum_{v \in C(P_{Si}^*)} w(v) = \sum_{i=0}^{K-1} \sum_{v \in C(P_{Si}^*)} \frac{|P_{Si}^*|}{|C(P_{Si}^*)|} = \sum_{i=0}^{K-1} |P_{Si}^*| = |\mathcal{R}|, \quad (2)$$

where $P_{S0}^* = \{S_0\}$, $|P_{S0}^*| = 1$, and $C(P_{S0}^*) = S_0$.

Suppose that $\mathcal{R}^* = \{S_1^*, ..., S_N^*\}$ is a minimum set cover for $\{V, \mathcal{S}\}$. Since an element of V may be contained in more than one subset of $\mathcal{R}^*$, it follows that

$$\sum_{v \in V} w(v) \leq \sum_{i=1}^{N} \sum_{v \in S_i^*} w(v). \quad (3)$$

Next we will show an inequality which bounds from above the total charge of a subset in $\mathcal{R}^*$, *i.e.*, for any $S^* \in \mathcal{R}^*$,

$$\sum_{v \in S^*} w(v) \leq D_C(G)(1 + H(|S^*| - 1)). \tag{4}$$

Let n_i $(i = 0, 1, ..., K)$ be the number of elements of S^* that have not been covered by $\mathcal{S}$ after iteration $i-1$, where step 1 is considered as iteration 0. Notice that $n_0 = |S^*|$ and $n_K = 0$. Let $\{i_1, ..., i_k\}$ denote the subsequence of $\{i = 0, 1, ..., K-1\}$ such that $n_i - n_{i+1} > 0$, *i.e.*, at iterations $i = i_1, ..., i_k$, at least one element of S^* is covered by P^*_{Si} for the first time. For each element v covered at iteration i_1, if $i_1 = 0$, based on the greedy rule at step 1, we have

$$w(v) = e(P^*_{S_0}) \leq \frac{1}{n_{i_1}}; \tag{5}$$

Otherwise, depending on whether a cover-adjacent subset or a graph-adjacent subset is selected at iteration i_1,

$$w(v) = e(P^*_{S_{i_1}}) = \left\{ \begin{array}{ll} \frac{|P^*_{S_{i_1}}|}{|C(P^*_{S_{i_1}})|} & \text{(cover-adjacent)} \\ \frac{1}{|C(P^*_{S_{i_1}})|} & \text{(graph-adjacent)} \end{array} \right\} \leq \frac{D_C(G)}{n_{i_1} - n_{(i_1+1)}}. \tag{6}$$

The inequality in (6) is due to three facts: (i) S_{i_1} is cover-adjacent with $\mathcal{R}$, leading to $|P^*_{S_{i_1}}| \leq D_C(G)$; (ii) $P^*_{S_{i_1}}$ covers at least $n_{i_1} - n_{(i_1+1)}$ elements of V, *i.e.*, $|C(P^*_{S_{i_1}})| \geq n_{i_1} - n_{(i_1+1)}$; (iii) $D_C(G) \geq 1$. Combining (5) and (6) yields

$$w(v) \leq \frac{D_C(G)}{n_{i_1} - n_{(i_1+1)}}. \tag{7}$$

The proof in [1] does not consider the case of $i_1 \neq 0$, leading to the wrong inequality

$$w(v) \leq \frac{1}{n_{i_1} - n_{(i_1+1)}}.$$

Consider two cases:

(i) If all the elements of S^* are covered after iteration i_1, *i.e.*, $n_{(i_1+1)} = 0$, then

$$\sum_{v \in S^*} w(v) \leq \sum_{v \in S^*} \frac{D_C(G)}{n_0} = D_C(G). \tag{8}$$

(ii) If not all the elements of S^* are covered by $\mathcal{R}$ after iteration i_1, S^* becomes cover-adjacent with $\mathcal{R}$ and thus a candidate for being selected at the following iterations. Then based on the greedy rule at step 2, we have that for an element $v \in S^*$ covered at iteration i_j $(j = 2, ..., k)$,

$$w(v) = e(P^*_{S_{i_j}}) \leq e(P_{S^*}) = \frac{|P_{S^*}|}{|C(P_{S^*})|} \leq \frac{D_C(G)}{n_{i_j}}. \tag{9}$$

Notice that if P_{S^*} is selected at iteration i_j, at least n_{i_j} elements will be covered for the first time, *i.e.*, $|C(P_{S^*})| \geq n_{i_j}$.

It follows from (7,9) that

$$\begin{aligned}\sum_{v\in S^*} w(v) &\leq (n_{i_1} - n_{(i_1+1)})\frac{D_C(G)}{n_{i_1} - n_{(i_1+1)}} + \sum_{j=2}^{k}(n_{i_j} - n_{(i_j+1)})\frac{D_C(G)}{n_{i_j}} \\ &= D_C(G)\left(1 + \sum_{j=2}^{k}\frac{n_{i_j} - n_{i_{(j+1)}}}{n_{i_j}}\right). \qquad (10)\end{aligned}$$

Here we have used the fact that $n_{(i_j+1)} = n_{i_{(j+1)}}$. It is because between iteration i_j and iteration $i_{(j+1)}$, no elements of S^* are covered.

For the summation term in (10), we have the following inequality:

$$\begin{aligned}\sum_{j=2}^{k}\frac{n_{i_j} - n_{i_{(j+1)}}}{n_{i_j}} &\leq \sum_{j=2}^{k}\frac{1}{n_{i_j}} + \frac{1}{n_{i_j}-1} + \cdots + \frac{1}{n_{i_{(j+1)}}+1} \\ &= H(n_{i_2}) \leq H(|S^*| - 1). \qquad (11)\end{aligned}$$

The last inequality is due to the fact that $n_{i_2} \leq n_{i_1} - 1 = |S^*| - 1$.

Eqn. (4) is a direct consequence of (8), (10), and (11). Thus, using (2-4),

$$\begin{aligned}|\mathcal{R}| &= \sum_{v\in V} w(v) \leq \sum_{i=1}^{N}\sum_{v\in S_i^*} w(v) \\ &\leq \sum_{i=1}^{N} D_C(G)(1 + H(|S_i^*| - 1)) \\ &\leq D_C(G)(1 + H(\gamma - 1))|\mathcal{R}^*|. \qquad \square\end{aligned}$$

Let $n = |V|$ be the number of elements of V. Then the approximation ratio of the modified greedy algorithm is $D_C(G)(1 + H(\gamma - 1)) = O(\ln n)$. Since the set cover problem is a special case of the MCSC problem where the auxiliary graph G is complete and the best possible approximation ratio for the set cover problem is $O(\ln n)$ (unless NP has slightly superpolynomial time algorithms) [2], the modified greedy algorithm achieves the order-optimal approximation ratio.

4. Connection with Connected Dominating Set Problem

A dominating set of a graph is a subset of vertices such that every vertex of the graph is either in the subset or a neighbor of some vertex in the subset. The connected dominating set (CDS) problem asks for a dominating set of minimum size where the subgraph induced by the vertices in the dominating set is connected. It is not difficult to show that the CDS problem is a special MCSC problem. Specifically, given an undirected graph $H = (V, E)$, we can derive an MCSC problem $(V, \mathcal{S}, G)$ from the CDS problem of H as follows:

(i) the universe set V is the vertex set V of H;

(ii) For each vertex $v \in V$, create a subset $S_v = \{v\} \cup \{\text{all neighbors of } v\}$ of V in $\mathcal{S}$;

(iii) the auxiliary graph G is the same as the given graph H except that each vertex of H is replaced by S_v, as illustrated in Fig. 1.

It can be shown that by exchanging the vertex subset S_v with the vertex v, the optimal solution to the derived MCSC problem is equivalent to the optimal solution to the CDS problem.

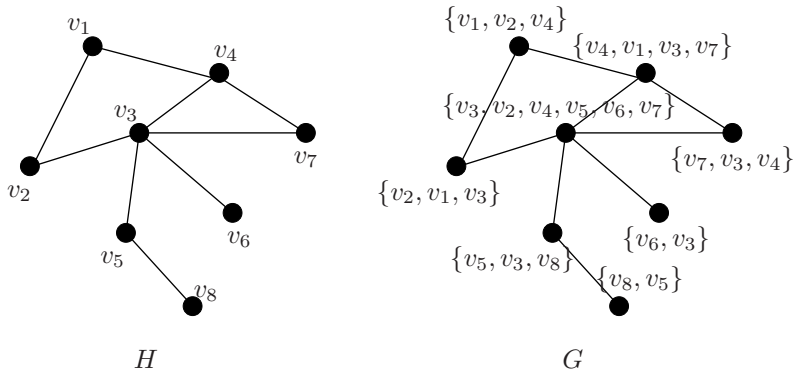

Figure 1: An illustration of the auxiliary graph G derived from the given graph H.

Guha and Khuller propose a greedy algorithm (Algorithm I in [3]) for CDS problem with an approximation ratio $2(1 + H(\gamma - 1))$, where $\gamma = \max\{|S_v| \mid v \in V\}$ and $\gamma - 1$ is the maximum degree of the vertices in H. The modified greedy algorithm for the MCSC problem reduces to the greedy algorithm of [3] when applied to the CDS problem. Notice that $D_C(G) = 2$ for the derived MCSC problem, since two vertex subsets S_{v1} and S_{v2} are overlapping if and only if their corresponding vertices v_1 and v_2 have at least one common neighbor. We see that the approximation ratio of the modified greedy algorithm established here is consistent with the one shown in [3], while the original approximation ratio obtained in [1] is not.

References

[1] Z. Zhang, X. F. Gao, W. L. Wu, Algorithms for Connected Set Cover Problem and Fault-Tolerant Connected Set Cover Problem, Theoretical Computer Science 410 (8-10) (2009) 812–817.

[2] U. Feige, A Threshold of $\ln n$ for Approximating Set Cover, Journal of the ACM 45 (4) (1998) 634–652.

[3] S. Guha, S. Khuller, Approximation Algorithms for Connected Dominating Sets, Algorithmica 20 (4) (1998) 374–387.